Contextual Object Detection with Multimodal Large Language Models

Year: 2023 | *Citations: 131* | *Authors: Yuhang Zang, Wei Li, Jun Han*

Abstract

Recent Multimodal Large Language Models (MLLMs) are remarkable in vision-language tasks, such as image captioning and question answering, but lack the essential perception ability, i.e., object detection. In this work, we address this limitation by introducing a novel research problem of contextual object detection—understanding visible objects within different human-AI interactive contexts. Three representative scenarios are investigated, including the language cloze test, visual captioning, and question answering. Moreover, we present ContextDET, a unified multimodal model that is capable of end-to-end differentiable modeling of visual-language contexts, so as to locate, identify, and associate visual objects with language inputs for human-AI interaction. Our ContextDET involves three key submodels: (i) a visual encoder for extracting visual representations, (ii) a pre-trained LLM for multimodal context decoding, and (iii) a visual decoder for predicting bounding boxes given contextual object words. The new generate-then-detect framework enables us to detect object words within human vocabulary. Extensive experiments show the advantages of ContextDET on our proposed CODE benchmark, open-vocabulary detection, and referring image segmentation.